Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1226 – 1231

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

Mining a Ubiquitous Time and Attendance Schema using Random Forests for Intrusion detection.

Promod KV[a*] Binu Jacob[a]

[a]*Cochin University of Science and Technology,Kerala,Cochin-22,India*
[a]*Cochin University of Science and Technology,Kerala,Cochin-22,*[*]*India*

Abstract

Our case study used Random Forest to measure the intrusion of unauthorized personnel to certain designated areas of the organization. Intrusions happening at high security areas could cost the organization in terms of loss of material and in worst cases of lives as well. The time attendance system also act as a security system as it involves access to doors and barriers through which only authorized personnel should access.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

Keywords: ubiquitous,Time and Attendance;Intrusion Detection

1. Introduction

Being a prominent airline of the Middle East, the use of a ubiquitous Time and Attendance (TA) application has been in place to manage time and attendance data of all staff. The application is interfaced with a set of clocks/readers, some of which are mounted with sensors that permeate the parking space and the entry doors of the building of the airline head quarters, hangars etc. Monitoring using CCTV is already present, but has its own limitations. For example the CCTV monitoring staff may not be present at any moment or security staff could overlook a suspect. In many cases of history, it was often camouflaged intruders with stolen ID cards or staff personnel who misused resources

* * Corresponding author. Tel.: +91 9895044162;
E-mail address: pramodvk4@gmail.com

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.097

that negatively impacted the security of an organization. Hence, security situation has at all times need to be vigilant in the areas involving the security of the aircraft itself , as any intrusion could lead to abnormal situations like hijack, explosions etc.

Ubiquitous Data Mining (UDM) is the process of performing analysis of data from ubiquitous computing environments. Our environment can be classified as a ubiquitous environment operating in a time-critical environment and includes continuous monitoring and analysis of status information received by readers at various locations for intrusion detection as well as time and attendance. Intrusion detection is a mechanism which helps identify the use of system for purposes other than intended.

This paper is organized as follows: Section 2 describes the related work in the field. Section 3 analyses our proposed route to solve this case study. Section 4 outlines the implementation and assumptions while Section 5 handles the experimentation results and analysis. Finally in Section 6 we make our concluding remarks.

2. Related Work

Mostly intrusion detection techniques are classified into misuse detection and anomaly detection. Anomaly detection helps to focus on detecting unusual activity patterns in collected data whereas misuse detection methods are intended to recognize known attack patterns. The method proposed in [1] Maintains an ensemble of intelligent Paradigms for Intrusion detection for building a network intrusion detection model that includes Support Vector Machines (SVM), ensemble voting system [2] ,[3] proposes an intrusion detection system using hybrid intelligent systems.

Performance of intrusion detection systems are comparatively studied in [4], application of Naïve Bayes to detect Network intrusion is studied in [5]. [6] Uses Support Vector Machines to study intrusion detection. In [7] the researchers studied how to identify a false alarm for network intrusion detection using hybrid data mining decision tree. A semi-Naïve Bayesian method for detecting of Network intrusions was proposed in [8]. Neuro-fuzzy techniques to reduce false alerts in intrusion detection was proposed in [9]. Classification using Random forests in data streams was proposed in [10]. Internet intrusion was studied in [11] and applied with a combination of Random Forests and Naïve Bayes' techniques. Combination of Support Vector Machines and Naïve Bayes' was proposed in [12] for Intrusion detection.

An intrusion detection system (IDS) is designed to monitor all inbound and outbound network activity and identify any suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. IDS is considered to be a passive-monitoring system, since the main function of an IDS product is to warn of suspicious activity taking place − not prevent them. The major difference between the above studies are that they are mostly intrusion detection in network environments whereas our study is about application of Random Forest in a data streaming environment [which is a Time and Attendance system and acts as a security application as well] for any intrusions, misuse etc. to physical locations of buildings. Hence, both are totally different though may share some underlying characteristics' and hence our study is one of the first such studies in an application environment.

3. Architecture, Database Infrastructure and Data Mining Model

The architecture in Figure1 shares more common features with the DAID (Database-centric Architecture for Intrusion Detection) architecture but differs in some areas. The card readers collected the data into a primary repository which was a Oracle database from where it was streamed in using an API (in real time) using a FTP tool and it's scheduler was routed to a folder from where the R software loads the data (acting as a middle ware) and classifies the data and thereafter sends into a GUI based java program .This java program was an application developed for the security department to understand the data in readable layman terms. The major benefits of using such an integrated approach are improved security, speed, data management and ease of implementation.

The database of our Time and Attendance application is hosted on a Sun server (Sun Fire V880, 32GB RAM, Solaris 2.8 OS) and the schema of application is implemented in Oracle 11g.

4. Explaining the Data and the working of Random Forests algorithm.

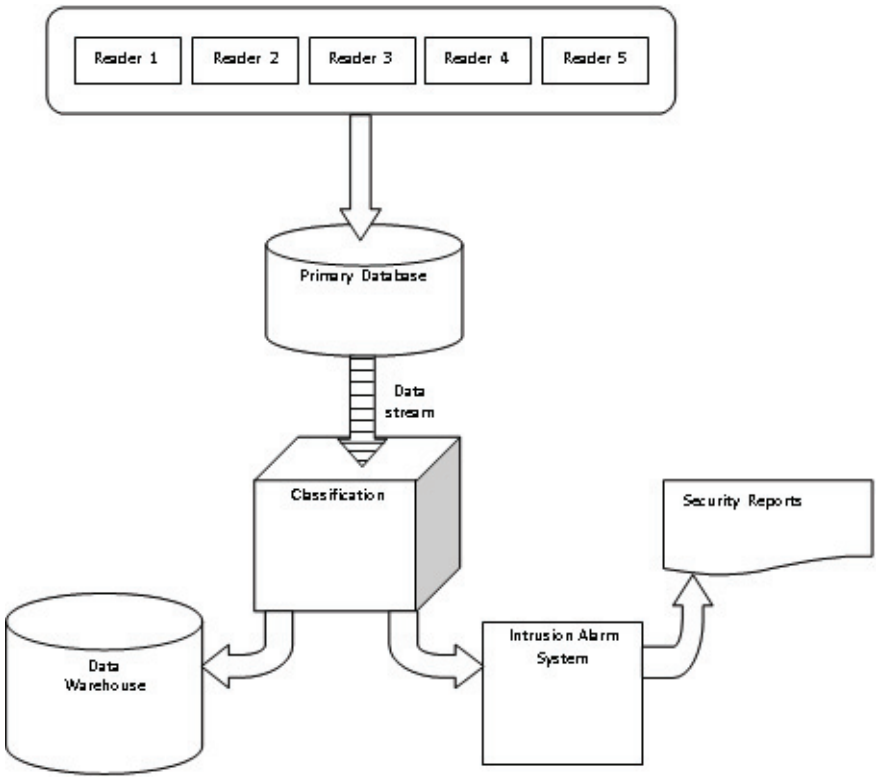

Fig.1. Integrated Mining model for intrusion detection in a TA schema

Figure 1 shows the model which was developed. Our aim was to understand and classify to the extent there were misuse in accessing resources of the readers by staff personnel using their unique legic advent cards of the organization which could hamper the security of the physical locations. This attempt was difficult as there often vague distinctions between errors punches and attempts to misuse the system. This could be only identified by identifying repeated attempts which could lead to a pattern by recognizing the time of the attempt, the sensitivity of the location. For this, we analyzed a data stream of three quarters of three different years of historical data and real time data for four months. Figure 2 shows the process.

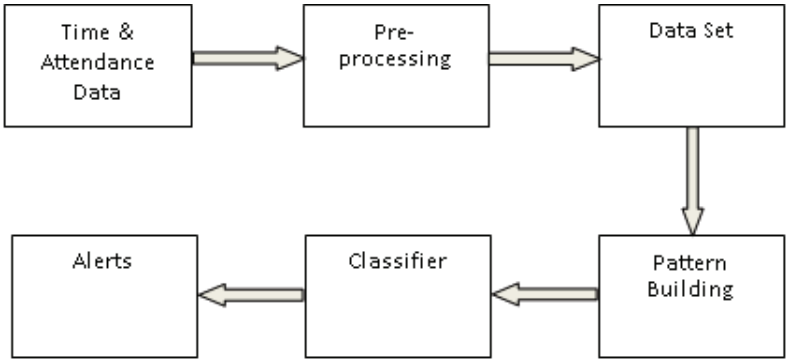

Fig..2.Process Model

We chose Random Forests as both data and independent variables are selected at random. This was well suited to the challenge, because the instances of punch at certain sensitive locations were also random .It also provides the class of dependent variables based on many trees. Independent variables were used to grow the tree which has the following characteristics (i) Most of the trees can make the right prediction of class for most part of the data (ii) Trees can only classify data wrongly in some instances. (iii) By using many decision trees we are able to classify the data better. In short, the choice of Random Forest is to generate multiple little trees from random subsets of data. Each of them is capturing different regularities since random subset of the instances are in the interest. In our study, we were able to obtain more cluttered decision boundaries than simple lines.

The attributes of the data were as follows: l =>location, t=> time, n=> name of card reader, a=> access or attendance, p=> authorized profile of staff, c=>intrusion or error punch or misuse. The target class attribute was ‘c’ which was independent of other attributes. We used the basic Streaming Random Forests algorithm which has classification accuracies approximately equal to those of the Standard Random Forest algorithm. Once a set (block) of records were chosen, the algorithm is followed .With each newly arrived record it was routed down the tree construction, based on its attribute values, until it reaches a node frontier, where the attribute values of the record are used to calculate ‘class counts’ that build up the computing of Gini indexes. The values of each attribute chosen (ordinal or numerical) are discretized into fixed length intervals. The split points are considered from the boundaries between the intervals.

The Hoeffding bound test has to be satisfied to transform the frontier node to an internal node with an inequality based on the best attribute and split points were given by the Gini index tests. Now the two children of this node becomes new frontier nodes .If a threshold is broken by the number of records that reached the frontier node, and the node has not been split, the algorithm transforms the node into a leaf, if the accumulated records are mostly all from the same class. If not, the node is transformed to an internal node which is always based on the split point and best attribute.

We used historical and real time data from the Time and Attendance repository in our experiments. The total number of records in both training and test we used were 28561.

5. Results and Analysis

The results showed that the algorithm achieves classification accuracies that exceeded our expectations. The uniformity in our data type as it streamed from the same system with fixed data types rather than a heterogeneous system helped us to keep the classification errors at minimum. This helped our multi class prediction to have better results in terms of accuracy even though the areas close to decision boundaries were harder to separate as they intersected. Another factor that has increased the quality of our classification result was the fact that we averaged the

training records from a period of three different years for historical data. This helped to show substantial correlation between the train and test set of data. The variables of importance also gave us indication as to what needs to be watched over by the security. The card reader name (n), location (l), time (t) and access or attendance (a) were the variable order priority. This implied that card readers at high sensitive locations at specific times needed additional security cover to ward off any potential threats.

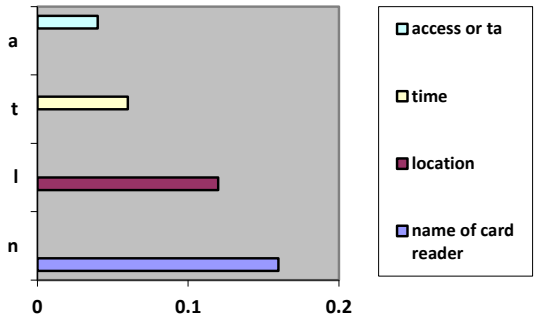

Fig.3. Mean Decrease Accuracy

Table.1. Classification results

		Intrusion Attempt (Predicted)	Error Punch (Predicted)	Misuse (Predicted)	Accuracy
Classification on TA schema	**Intrusion Attempts (Actual)**	**637**	**62**	**12**	**90.1%**
	Error Punch (Actual)	**1**	**1305**	**5**	**98.9%**
	Misuse (Actual)	**1**	**7**	**1073**	**99.1%**

6. Conclusion

This case study is different from other studies as Random Forests was applied in a Ubiquitous Time and Attendance schema application which gave us fairly good classification results. This study has helped improve the security situation at physical locations of the Company. Further work can be improved upon by including Concept drift and Novel class challenges as well.

References

[1]S.Mukkamala,A.H. sung and Ajith Abraham. Intrusion detection using an ensemble of intelligent paradigms. Journal of Network and Computer Applications. Vol.28.pp. 167-182,2005

[2]M.Panda and M.R.Patra.Ensemble voting system for anomaly based network intrusion detection. International Journal of recent trends in Engineering,Vol.2,No.5,pp. 8-13,2009

[3]S.Peddabachigari.A.Abraham,C.Grosan and J.Thomas. Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications. Vol.30.pp.114-132,2007.

[4]V.Venkatechalam and S.Selvan,Performance Comparison of intrusion detection system classification using various feature reduction techniques. International Journal of Simulation. Vol.9.No.1, pp 30-39,2008.

[5]M.Panda and M.R.Patra, Network Intrusion Detection using Naïve Bayes. International Journal of Computer Science and Security, Vol.7,No 12,pp. 258-263,2007.

[6]Sung-Hae Jun and Kyung-Whan oh, An evolutionary support vector machine for intrusion detection. Asian Journal of Information Technology,Vol.5,No.7,pp. 778-783,2006.

[7]N.B Annur,H.Sallehudin,A.Gani and O.zakari.Identifying false alarm for Network Intrusion detection system using hybrid data mining decision tree. Malaysian Journal of Computer Science,Vol.21, No. 2,pp. 101-115,2008.

[8]M.Panda and M.R.Patra. A semi-Naïve Bayesian method for detecting network Intrusions.LNCS,Vol.5863,pp.614-621,2009.

[9] P.Gaonjur,N.Z.Tarapore and S.G.Pokale using neuro-fuzzy techniques to reduce false alerts in intrusion detection. In.Proceedings of International Conference on Computer Networks and Security,India,pp. 1-6,2008.IEEE Press.

[10] Hanadi Abdul Salam ,D.B.Skillicorn and P.Martin: Classification Using Streaming Random Forest, Published by the IEEE Computer Society,2011

[11]Younes Chihab,Abdelah,Mohammed Erritali and Bouabid: Detection & Classification of Internet Intrusion Based on the Combination of Random Forest and Naïve Bayes, International Journal of Engineering and Technology,Jun'2013

[12]Aman Mudgal and Rajiv Munjal: Using Support Vector Machine and Naïve Bayes Classification for Intrusion Detection, International Journal for Innovative Research in Science and Technology,Feb' 2015.